Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 45-55 recibido 13-12-2023 revisado 15-01-2024 aceptado 29-01-2024

An Empirical Study on the Number of Items in Human Evaluation of Automatically Generated Texts

Estudio Empírico sobre el Número de Elementos en la Evaluación Humana de Textos Generados Automáticamente

Javier González-Corbelle,[1] **Jose M. Alonso-Moral,**[1] **Rosa M. Crujeiras,**[2] **Alberto Bugarín-Diz**[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
[2]Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga),
Universidade de Santiago de Compostela, Spain
{j.gonzalez.corbelle, josemaria.alonso.moral, rosa.crujeiras, alberto.bugarin.diz}@usc.es

Abstract: Human evaluation of neural models in Natural Language Generation (NLG) requires a careful experimental design in terms of the number of evaluators, number of items to assess, number of quality criteria, among other factors, for the sake of reproducibility as well as for ensuring that significant conclusions are drawn. Although there are some generic recommendations on how to proceed, there is not an established or accepted evaluation protocol admitted worldwide yet. In this paper, we address empirically the impact of the number of items to assess in the context of human evaluation of NLG systems. We first apply resampling methods to simulate the evaluation of different sets of items by each evaluator. Then, we compare the results obtained by evaluating only a limited set of items with those obtained by evaluating all outputs of the system for a given test set. Empirical findings validate the research hypothesis: well-known resampling statistical methods can contribute to getting significant results even with a small number of items to be evaluated by each evaluator.
Keywords: Natural Language Generation, Human Evaluation, Resampling Methods.

Resumen: La evaluación humana de modelos neuronales en Generación de Lenguaje Natural (GLN) requiere un diseño experimental cuidadoso de elementos como, por ejemplo, número de evaluadores, número de ítems a evaluar, número de criterios de calidad, entre otros, para así garantizar la reproducibilidad de experimentos, así como para asegurar que las conclusiones extraídas son significativas. Aunque existen algunas recomendaciones genéricas sobre cómo proceder, no existe un protocolo de evaluación consensuado, general y aceptado. En este artículo prestamos atención a cómo influye el número de elementos a evaluar en la evaluación humana de los sistemas de GLN. Aplicamos distintos métodos de remuestreo para simular la evaluación de distintos conjuntos de ítems por parte de cada evaluador. A continuación, comparamos los resultados obtenidos evaluando sólo un conjunto limitado de ítems con los obtenidos evaluando todas las salidas del sistema para el conjunto completo de casos de prueba. Las conclusiones derivadas del estudio empírico corroboran la hipótesis de investigación de partida: el uso de técnicas de remuestreo ayuda a obtener resultados de evaluación significativos incluso con un número pequeño de ítems a evaluar por cada evaluador.
Palabras clave: Generación de Lenguaje Natural, Evaluación Humana, Remuestreo.

1 Introduction

There is debate about the use of automatic metrics versus human judgement when evaluating the output of Natural Language Generation (NLG) systems. Reiter (2018) stated that commonly used automatic metrics, such as ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), or BLEU (Papineni et al., 2002), do not correlate well with human judgements for the evaluation of NLG systems. This is mainly because the most popular metrics are based on checking the n-gram overlap of the generated sentence with a limited set of reference texts that are considered correct, but do not cover all the possible text variations that NLG systems may

ISSN 1135-5948 DOI 10.26342/2024-72-3

produce (e.g., paraphrases, synonyms, or alternate realizations). Accordingly, other metrics have emerged, such us embeddings-based metric to measure similarity between reference and candidate texts like BERTScore Zhang et al. (2020), or pre-trained metrics, i.e., neural models trained to learn how to automatically do an evaluation task, like BLEURT (Sellam, Das, and Parikh, 2020) or NUBIA (Kane et al., 2020). More recently, there are studies regarding the application of ChatGPT for assessing generation tasks (Wang et al., 2023). However, despite these efforts to produce more and more data-driven automatic metrics, which are inspired from the machine learning community, the lack of correlation with human evaluation persists (Moramarco et al., 2022).

On the other hand, Van der Lee et al. (2021) recommended some best practices for human evaluation, and the NLG research community is doing efforts to set the basis for reproducible human evaluation (Belz et al., 2023; Belz, 2022). But, in spite of this, there is still a lack of formal protocol for carrying out NLG human evaluation. Furthermore, conducting human evaluation properly is not straightforward, since there are multiple factors that must be considered, being among them the textual properties to be assessed, the evaluation criteria that human evaluators must follow, the number of human evaluators, the number of items to evaluate, the number of questions per item, the statistical tests, tools for data analysis, etc.

In this paper we focus on validating the following research hypothesis: "well-known resampling statistical methods can contribute to getting significant results even with a small number of items to be evaluated by each evaluator". Thus, we aim to prove empirically the influence of the number of items presented to an evaluator in the context of human NLG evaluation. Starting from a set of texts generated by an NLG system (i.e., a set of items to be evaluated), we research on the minimal number of texts to be assessed for ensuring that the evaluation results obtained are significant.

More precisely, we apply two resampling methods to simulate multiple evaluations, thus exploring the effect of different number of items per evaluation. As far as we know, this is the first empirical study regarding the impact of the number of items to assess in NLG human evaluation. Notice that the concept of "item" may vary depending on the context. In the context of NLG evaluation, some researchers may understand as "item" each criteria used to manually evaluate the text (e.g., coherence, quality, etc.), but in this paper item refers to each text to be evaluated.

The rest of the manuscript is organized as follows. Section 2 introduces some preliminary concepts. Section 3 presents the methods to be used for the experimentation described in Section 4. Finally, Section 5 concludes the paper with some final remarks and points out future work.

2 Background

One of the parameters to be set when designing a human evaluation process is the number of items (i.e., either the number of questions an evaluator must answer or the number of tasks an evaluator must do) to obtain sufficiently reliable and representative results, while avoiding work overload. However, selecting a representative number of items is not a trivial task and depends on the type of study you are conducting.

In the field of statistics, there was a tendency to use the "n=30 rule-of-thumb" and set at least 30 as the default minimal number of questions or tasks for any study, but, to the best of our knowledge, without any scientific justification or empirical evidence. A possible explanation for this may have its origin in the pre-computer era, when all the calculations were made by hand. Student (1908) described how, when calculating the probable error of correlation coefficients, the best results were obtained with a sample size of 30 and one of the conclusions was "with samples of 30 [...] shows that the mean value approaches the real value [of the population] comparatively rapidly". Afterwards, the choice of this "magical" number as a sufficient sample size to get sounded results (from a statistical viewpoint) was maintained for decades, arguing that 30 samples were enough to hold the central limit theorem. Years later, in the computer era, this belief was deprecated in favor of bootstrap-based diagnostics (Hesterberg, 2008).

In the context of NLG human evaluation, Van der Lee et al. (2021) stated that "there should also be a sufficient number of outputs, so that a couple of particularly good or bad items do not skew the results too much. However, the number of items to evaluate depends heavily on the diversity of the sample, so we cannot give any specific recommendations here.". In their analysis of 89 papers, they noted that there was a median of 100 items utilized for human evaluation. However, the quantity of items varied widely, ranging from 2 to 5400, indicating a significant disparity. Thus, there is no general rule to determine the

number of items that must be evaluated to obtain a reliable evaluation of an NLG system. Of course, the smaller the number of items required to get significant results the better. But beware of the negative oversimplification risk.

3 Methods

We aim to measure empirically the influence of the number of items when carrying out human evaluation of an NLG system. Considering that the number of items in our context refers to the number of texts to be evaluated when assessing the goodness of an NLG system, we will start from a large pool of texts evaluated by humans (what is taken as the baseline). Then, we will apply different resampling strategies in the search for the minimal set of texts taken from the initial pool that is required to draw sounded conclusions, i.e., to extract insights with a reasonable statistical significance. Before going in depth with the experimental study, let us introduce the resampling methods (see Section 3.1) and statistical tests (see Section 3.2) to be used later in Section 4.

3.1 Resampling methods

In statistics, resampling refers to the creation of new synthetic samples from observed or real ones. There are different methods to perform resampling such as cross-validation, permutation, subsampling or bootstrap. The latter is the one we used in our experiments and is introduced here to better understand the upcoming sections.

Bootstrap is an approach to statistical inference proposed by Efron (1979) which translates, in practice, the construction of different resampling schemes to approximate the sample distribution of a statistic (i.e., a function of the sample). The basic idea of bootstrap is that it is possible to make inferences about a given population from a reduced but representative sample of such population. If the target population were known, then we may measure the degree of agreement between the data distributions associated with the selected sample and with the entire population. Imagine that we desired to use answers to a survey to predict the result of a coming election in a city. If we may conduct the survey with all people who is entitled to vote, then we may have a high confidence in the predictive power of results of such survey. However, conducting a survey with the entire population is very expensive and sometimes even impossible because some people may refuse to take part. In practice, we should look for the smallest sample of the population that is representative enough of the entire population, so we can optimize resources and maximize the chance of getting significant results. However, what is the size of the smallest (but yet representative, for a certain significance) sample? Can we say that collecting answers to the survey by 30 people is enough? There is not a magical number a priori for the optimal sample size because the predictive power of the survey is not only a matter of quantity but also a matter of "inference" quality.

With the bootstrap method, the inferences are performed regarding synthetic samples, and they are derived by resampling from the given data which represents a subset of the target population. The true error in a sample statistic against its population is unknown because the entire population is unknown. However, the quality of inference of each synthetic sample from resampled data is measurable if we take as baseline the full initial sample (assuming it represents well the data distribution in the entire target population).

In short, the procedure to apply bootstrap is as follows:

1. Obtain a data sample that will be the "population" over which subsampling is applied.
2. Choose the number of synthetic subsamples ($replications$) to be generated.
3. Choose a subsample size (s) per $replication$.
4. For each $replication$:
   (a) Produce a new synthetic subsample with replacement of size s.
   (b) Estimate the quality of the generated subsample by computing the desired statistic.
5. Aggregate statistics for all $replications$.

It is worth noting that using the bootstrap method, for each replication, we always get a sample of the chosen size with replacement. For example, if the population includes 4 values such as [a,b,c,d] and we set 3 as subsample size, then we may obtain something like [a,a,c], where some values in the generated sample can be repeated. Thus, all values in the original population have the same probability to be selected when filling in each position in each new synthetic sample. On the contrary, if we applied Resampling without Replacement (RWOR), for each replication, then repetition of values is not allowed. In our experiment, we will test both bootstrap and RWOR.

3.2 ANOVA for discrete distributions

The analysis of variance (ANOVA) is a statistical test to compare the means of two different groups or populations (Fisher, 1992). ANOVA produces as output a number (named as F-statistic) and a p-value which supports or rejects the null hypothesis. In an ANOVA test, the null hypothesis is that the means of the groups being compared are the same, while the alternative hypothesis is that group means are different. This way, if the p-value obtained from the test is less than the usual α significance levels (0.1, 0.05, 0.01), then the null hypothesis can be rejected, and we can state that at least one of the means is different from the others. Then, different post-hoc tests can be applied to find out for which specific group the mean is different Otherwise, the null hypothesis cannot be rejected and therefore, we do not have enough evidence to say that there is a significant difference between the groups under comparison.

Even if ANOVA was originally defined for continuous data, there are some ANOVA extensions to treat properly also discrete data (De Leon and Zhu, 2008). In our case, we will use a variation of the ANOVA test which is more suitable to deal with discrete distributions. Namely, the function used is called `discANOVA`, from the WRS2 package in R (Mair and Wilcox, 2020). This function checks if the null hypothesis (i.e., that for two or more independent groups, the corresponding discrete distributions are identical) is satisfied. More precisely, `discANOVA` verifies if the groups have identical multinomial distributions.

It is worth noting that the power analysis done with software tools like G*Power (Faul et al., 2009) helps designers estimate the number of participants that are required in a user study with the aim of achieving significant results. However, as far as we know there is not any power analysis associated with the estimation of the number of items to assess by each participant.

4 Experimentation

For testing empirically the influence of the number of items in an NLG evaluation procedure, we followed the next steps: (i) we used a real NLG system to generate some texts (see Section 4.1); (ii) we proceeded with the human evaluation of all the generated texts (see Section 4.2); (iii) we created different prototypical evaluator profiles (see Section 4.3); and (iv) we tested the influence of the number of items for each of the evaluator profiles previously defined (see Sections 4.4 and 4.5).

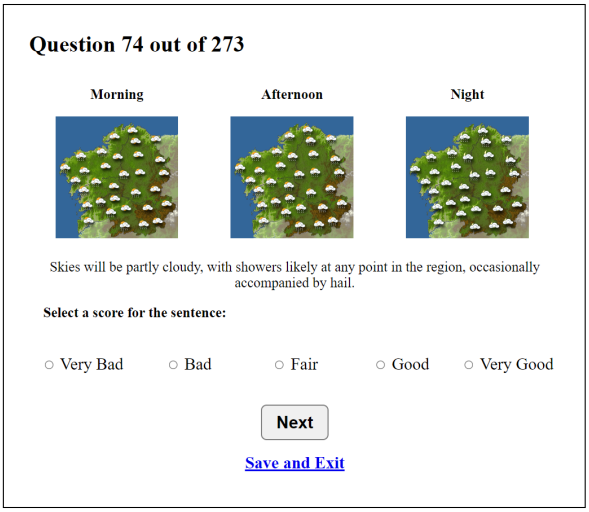


Figure 1: Example of question in the survey.

4.1 NLG system

We searched in the literature for a neural NLG system that may be used for generating the pool of texts to be evaluated, and we found that González Corbelle et al. (2022) released an NLG system along with the related dataset. Moreover, all the resources explained in the paper were available to reproduce the generation of texts.

The NLG system consists of a neural model which is adapted for a data to text (D2T) task in the context of meteorology. It generates short textual descriptions from meteorological tabular data. More precisely, it is an adaptation of a Transformer-based D2T model that initially was designed to generate chart captions (Obeid and Hoque, 2020). Regarding the dataset, we used the same as in the original model, composed of more than 3000 pairs of meteorological data and texts. We trained our model from scratch following the instructions given in the original paper (regarding the same parameters as well as the training, validation, and test partitions). As a result, we produced a total of 273 texts associated with the given test partition. These texts constitute the population to be evaluated in the following sections.

4.2 Human Evaluation

We designed an evaluation survey for assessing all the automatically generated texts. Each question from the survey was composed with a representation of the tabular input given to the system and the generated text.

The tabular data inputs represent the state of the sky for 32 different meteorological zones and 3 periods of the day (morning, afternoon, and night). Interpreting these 96-values table and

Score levels for evaluation	
Very bad	The description is not readable and does not match the data shown in the images, hallucinations are perceived in the generated text.
Bad	The description is not easy to read even though the content of the text is correct, but it ignores important information.
Fair	The description is readable, but not excessively natural. What is mentioned in the text is present in the data, but it is not complete enough.
Good	The description is well constructed, readable, and natural, but perhaps it could have mentioned some other relevant data present in the images.
Very Good	The description is so readable, natural, complete, and consistent with the data shown that could be considered a human text.

Table 1: Instructions given to the evaluators to score the texts based on their fluidity, naturalness, and content.

checking if the generated text describes the data correctly is a tedious task for an evaluator, so we opted for a simplified view of the input data. More precisely, we used the images available in the original data repository[1] instead of providing evaluators with raw data. For each question in the survey (see example in Figure 1), evaluators had to look at 3 meteorological maps (i.e., one for each period of the day) and rate how well (in a 5-point Likert scale from "Very Bad" to "Very Good") the observed state of the sky is described by the given text (which was automatically generated by the D2T system). Before the evaluation, evaluators were given clear instructions about how to score the texts, based on their fluidity, correctness, and content (see Table 1).

[1] https://gitlab.citius.usc.es/gsi-nlg/meteogalicia-es

Annotators	Cohen's κ	Fleiss' κ
1 vs. 2	0.2128	0.2188
1 vs. 3	0.2565	
2 vs. 3	0.2119	

Table 2: Inter-Annotator Agreement coefficients: pair-wise Cohen's Kappa and global Fleiss' Kappa.

Three different evaluators with experience in the NLG field assessed the 273 texts generated by the system. Their Inter-Annotator Agreement was calculated using both the Cohen (1960) and Fleiss (1971) Kappa coefficients (see Table 2). Regarding the pair-wise agreement, i.e., the Cohen's Kappa coefficient, we observe how the degree of agreement among the pairs of evaluators is similar in general (between 0.2 and 0.3). Nevertheless, going more into detail we could conclude that annotators 1 and 3 have the best agreement on their responses. Regarding the global agreement, i.e., the Fleiss' Kappa coefficient, the reported value (0.2188) is in the same range of values reported by the Cohen's coefficient. According to the Kappa statistic interpretation (Altman, 1991), both the pair-wise and global agreements are in the range $0.21 - 0.4$, that it is considered a "Fair Agreement". However, the coefficients obtained are closer to the low part of the range, especially the Fleiss' Kappa, which is far from the $0.41 - 0.6$ range considered as "Moderate Agreement" and even further from the "Good/Substantial Agreement" $(0.61 - 0.8)$ that is deemed as desirable. This highlights the difficulty of the NLG evaluation task.

4.3 Evaluator profiles

For the sake of generality, we designed five different prototypical evaluator profiles taking as reference the responses collected in the previous survey. This was done in this way because we were looking for synthetic but realistic prototypical profiles.

On the one hand, evaluators with tendency to score high ("Good" or "Very Good") most texts and not to penalize too much bad texts are considered to belong to a positive profile, while the evaluators whose tendency is just the opposite, i.e., to rate low ("Very Bad" or "Bad") most texts, are considered as belonging to a negative profile. On the other hand, we can consider "bipo-

lar" evaluators who tend to both extremes, i.e., only use very high and very low scores, or "neutral" evaluators who tend to "Fair" score for most cases. Bipolar evaluators are associated with a polarized profile, while neutral evaluators are associated with a neutral profile. In addition, there is a random profile which represents evaluators who vary randomly their scores in each question without any pre-defined criteria and do not fit in a specific evaluator profile of those already defined.

Considering these five evaluator profiles (positive, negative, neutral, polarized, and random), we proceed to generate their characteristic score distributions, by simulating as if an evaluator belonging to each of the profiles had evaluated all the cases under study. We take as a starting point the real scores collected in the previous survey and the generation procedure is made up of the following three steps:

1. **Transform all the responses from the three real human evaluators into three categories**: *negative*, *fair*, and *positive* responses.[2] Since we had five possible scores in a 5-point Likert scale, we aggregated all responses corresponding to "Very Bad" or "Bad" in the "negative" category, while responses associated to "Good" or "Very Good" go to the "positive" category. The "fair" category is made up of all responses with "Fair" scores.

2. **Aggregate the three evaluators' scores into a global curated score**: For each question, we apply the majority voting aggregation rule, i.e., if there are at least two evaluators that agree in the category of the response (negative, fair, positive), then such category defines the global score. Otherwise, the question is discarded from the aggregation process. As a result of the aggregation stage, we have a dataset with 246 curated cases, i.e., those cases in which there is at least a two-evaluators agreement.[3] The global distribution of scores is as follows: 89 cases are negative, 35 cases are fair, and 122 cases are positive (see Figure 2d).

3. **Create the five prototypical evaluator profiles**: We re-define 5-value distributions following the evaluation tendencies that are characteristic for each prototypical profile. From the dataset that we curated in the previous step, depending on the selected profile, we re-assign the cases into a different percentage for "Very Bad", "Bad", "Fair", "Good", and "Very Good" scores. Figure 2 shows the resultant score distribution for each evaluator profile, based on the profiles of the three human evaluators that were taken as reference. Figure 2d depicts the aggregated global scores from which the different evaluator profiles were generated. To do that, for each profile, the negative values are reassigned as "Very Bad" and "Bad" scores, while the positive values are reassigned as "Good" and "Very Good" scores. All the transformations are made in agreement with the expected tendencies for each evaluator profile. For example, if we look carefully at the distribution of cases in the Positive Profile (see Figure 2i) we can notice that from the 89 negative aggregated scores (see Figure 2d) 5% of cases are associated to the "Very Bad" score and 95% of cases are associated to the "Bad" score. Regarding the 122 positive values in the same picture, 50% of cases are associated to the "Good" score and the other 50% of cases is associated to the "Very Good" score. This way we produce a synthetic distribution of scores which is realistic (because it is grounded on the original human evaluations) but follows the expected tendency towards positive optimistic scores. Similar transformations produce the rest of profiles as depicted from Figure 2e to Figure 2h, always taking as starting point the aggregated global scores (Figure 2d) extracted from the real human evaluations.

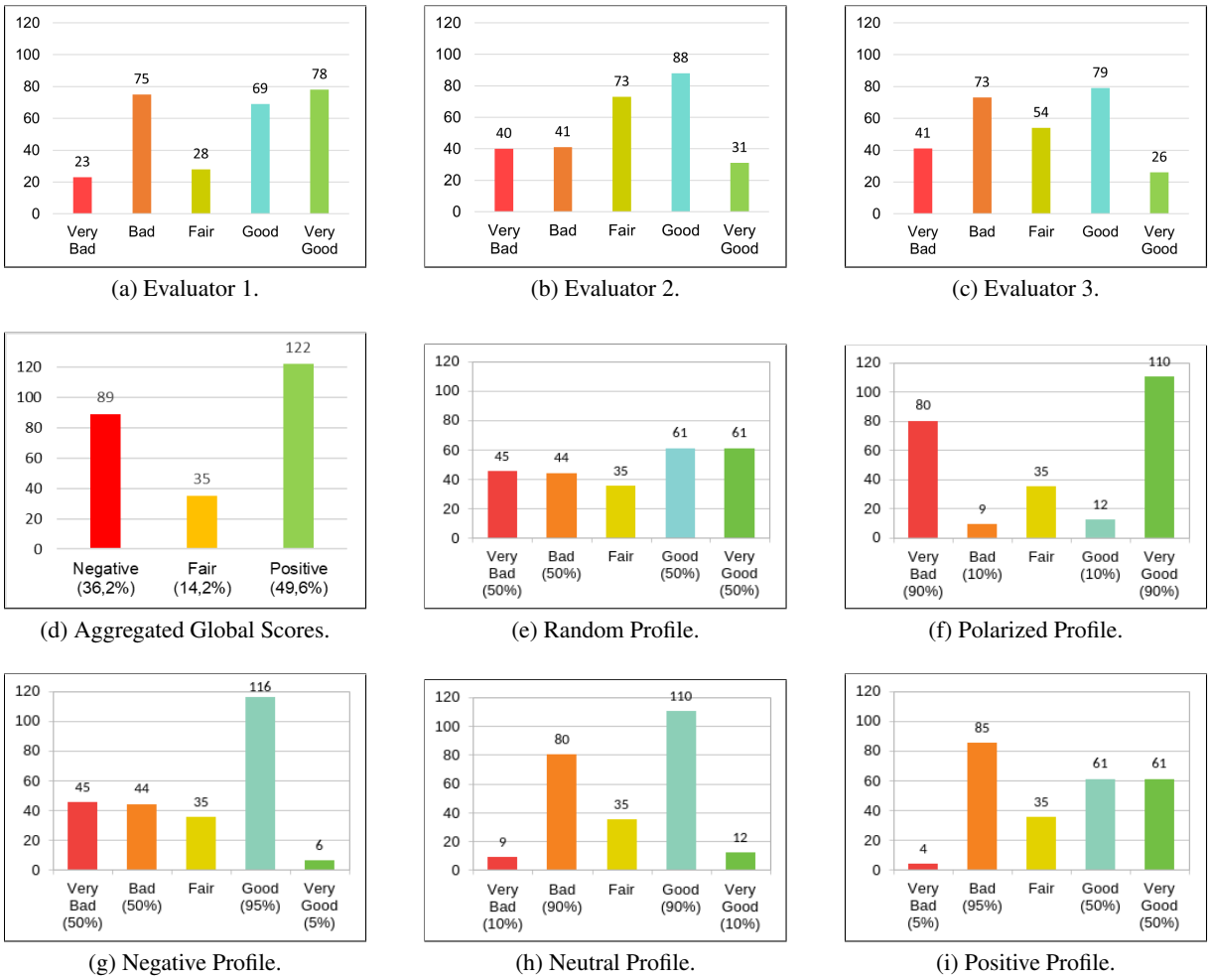
[2]Note that here we are talking about response categories and not about evaluator profiles.

[3]Due to disagreement among evaluators, 27 cases were discarded.

4.4 Resampling tests

In this section, we describe how to test the influence of the number of items in an NLG evaluation procedure. Considering the whole set of texts (S) generated by a system for a given test partition, if we evaluate a subset of texts $\tilde{S} \subset S$, some questions arise:

- How representative is $\tilde{S}$ (with respect to S)?
- Is the score distribution when evaluating $\tilde{S}$ the same as when evaluating S (considering the same evaluator or at least the same prototypical evaluator profile)?
- Which is the minimum number of samples in $\tilde{S}$ for yielding results as precise as the



(a) Evaluator 1. (b) Evaluator 2. (c) Evaluator 3.

(d) Aggregated Global Scores. (e) Random Profile. (f) Polarized Profile.

(g) Negative Profile. (h) Neutral Profile. (i) Positive Profile.

Figure 2: Original scores set by human evaluators (2a-2c), overall aggregated scores (2d), and generated score distributions for the synthetic prototypical evaluator profiles (2e-2i). The number below the bar labels is the percentage of negative/positive aggregated scores that were transformed into each level.

ones obtained with the whole set S?

- Does the number of items selected for $\tilde{S}$ have the same influence on different evaluators (or at least on different prototypical evaluator profiles)?

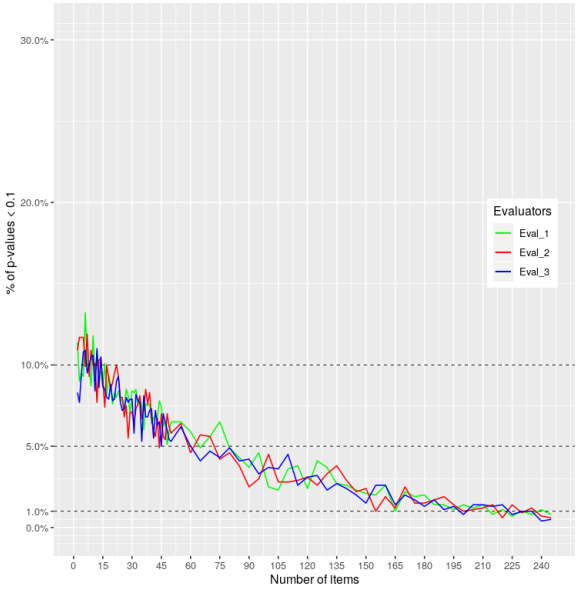
In the search for answers to the previous questions, the research hypothesis to validate is the following: we can approximate the "real" score distribution of an evaluator (i.e., the score obtained when such evaluator evaluates all texts in S) by evaluating only the items in $\tilde{S}$ and then applying a resampling method.

With the aim of testing if we can accept/reject the previous hypothesis, we apply bootstrap and RWOR resampling methods (as described in Section 3.1) on all the distributions shown in Figure 2. We set $\alpha = 0.1$. The number of replications is set to 1000. The sample size, which corresponds to the number of items to evaluate, ranges from 2 to 245. For each number of items tested in each of the distributions, we get 1000 p-values from the `discANOVA` output (i.e., one per replication). For each p-value, if it is lower than α, then we can say that in the given replication the resampled set of items $\tilde{S}$ has a distribution deemed as statistically different from the entire population S. It is worth noting that we count how many replications (out of 1000) yield to reject the null hypothesis: "the means of the groups S and $\tilde{S}$ are the same".

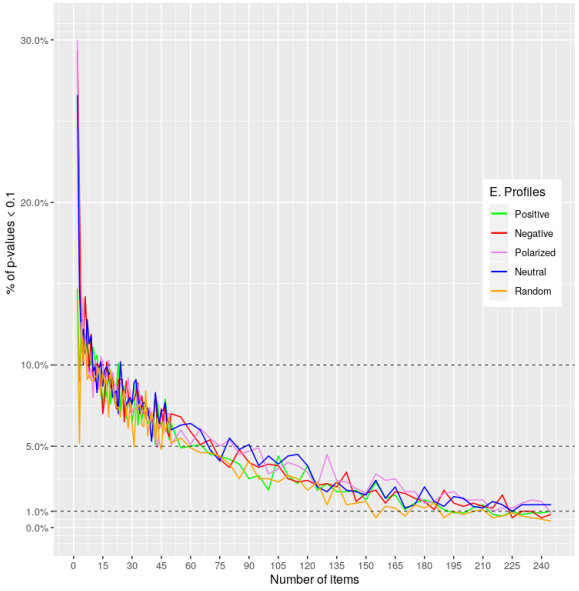
4.5 Results

Figure 3 summarizes the reported results. The comparison pays attention to the resampling methods (i.e., bootstrap and RWOR) and the type of evaluators (i.e., synthetic prototypical evaluator profiles vs. real evaluators).

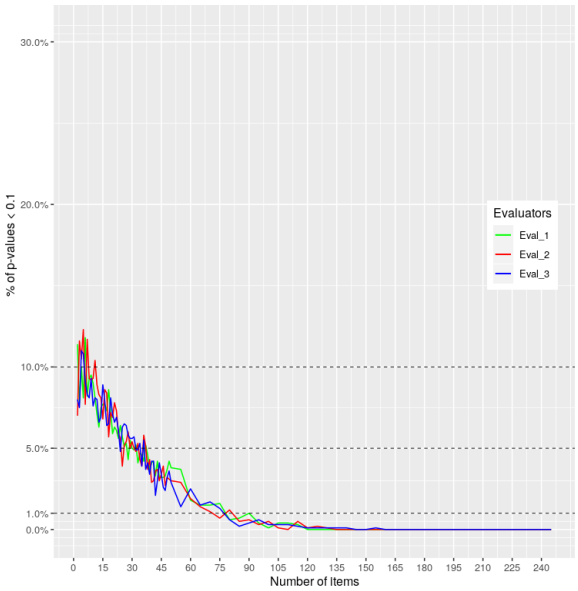
The general trend is that for a small number of items (i.e., less than 30) the hypothesis that



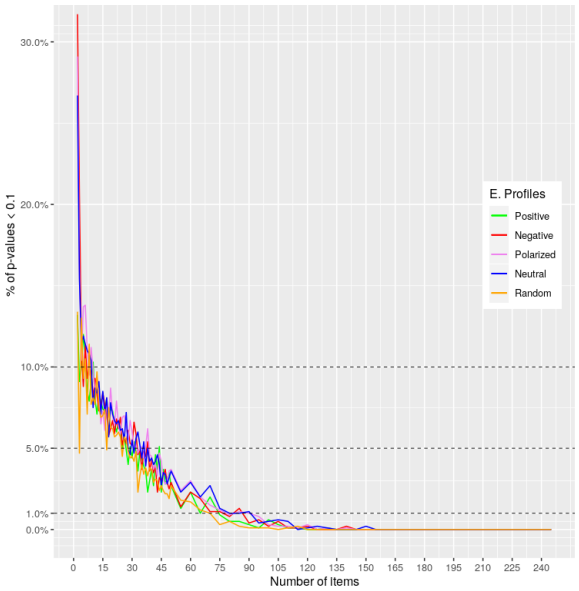
(a) Bootstrap: human evaluators.



(b) Bootstrap: prototypical evaluators.



(c) RWOR: human evaluators.



(d) RWOR: prototypical evaluators.

Figure 3: Results of running bootstrap and RWOR. For each number of items in the horizontal axis, 1000 samples were generated, and the picture shows the percentage of samples for which p-value < 0.1 when compared to the real distribution using `discANOVA`.

the two distributions are considered equal is rejected more often, no matter either the resampling method or the type of user. In the case of RWOR, values of 0% of hypothesis rejection are reached from 120 items on in some of the prototypical profiles (see Figure 3d), while for bootstrap the values tend asymptotically to 0% but never reach this value (see Figure 3b).

If we stick to reality, getting an exactly equal distribution from a small number of items is practically impossible. Therefore, we have established as a threshold of acceptance the cases in which less than 10% of the replications reject the hypothesis that the distributions are considered equal. Taking this threshold as a reference, we can see in Table 3 from what number of items less than 10% of cases reject the hypothesis, for each method and evaluator profile. The table also includes the minimal number of required items in case of establishing a smaller threshold value such as 5% or 1%. It is easy to appreciate how the smaller the pre-defined threshold, the bigger

the number of items that are required to get significant results.

	Bootstrap			RWOR		
Threshold	***10%***	***5%***	***1%***	***10%***	***5%***	***1%***
Eval_1	16	75	240	6	37	75
Eval_2	22	70	235	11	37	80
Eval_3	14	55	220	5	36	75
Positive	23	65	210	10	44	70
Negative	17	70	220	8	38	85
Polarized	18	80	240	9	38	85
Neutral	24	90	245	9	36	90
Random	18	55	210	8	27	65

Table 3: For each human evaluator and for each prototypical evaluator profile, number of items from which the % of samples with a p-value $<$ 0.1 is always lower than a pre-defined threshold as illustrated in Figure 3 (10%, 5%, 1%).

If we look at the bootstrap method, the evaluator that first reaches the threshold of 10% of rejection is the third one (Eval_3 in Figure 3a) with 14 items. In addition, the prototypical profiles that need the most items to achieve a distribution equivalent to the original one are the Neutral and Positive profiles, with 24 and 23 items, respectively (see Figure 3b). This is not the case for RWOR (see Figure 3c) which yields much lower numbers, with the distribution of Eval_3 corresponding to the lowest value (5 items), while Eval_2 is associated with the highest value (11 items).

On the one hand, taking the bootstrap method as a reference, with a rejection threshold of 10%, we could say that for any of the tested evaluators and prototypical profiles, from at least 24 items, we could obtain a distribution of scores equivalent to performing the complete evaluation of all the curated 246 test cases. On the other hand, if we consider the RWOR method for the same threshold, the number of items to obtain a distribution equivalent to the original one is reduced to no more than 11, for all evaluators and prototypical profiles tested in this experiment.

To sum up with, reported results validate our research hypothesis for the specific experimental setting under consideration, i.e., thanks to the use of resampling methods we can get sounded evaluation insights while requiring a very small number of items to be evaluated. The actual numbers are detailed in Table 3. For a 10% threshold, in the worst case, only 9.75% of the items in S needs to be evaluated. In the best case, the percentage of items to evaluate is only 2%. If a smaller threshold were required, then the number of items should be bigger. Moreover, the number of items required by RWOR is always much smaller. In the worst case, (i.e., the Neutral prototypical profile with threshold value equal to 1%) the required number of items is 36.58% of all the items in S.

5 Final Remarks and Future Work

In this paper we tested the influence of the number of items in a human evaluation of NLG systems. To do so, we first carried out an evaluation with three different raters on a pool of texts generated by a Data-To-Text neural system. Then, with the scores obtained from the evaluation of all the texts, we created different prototypical evaluator profiles (that are synthetic but realistic because they are grounded on the previous human evaluations). Finally, using resampling methods, we simulated evaluations in the search for the minimal number of items that is required to get sounded insights.

After carrying out the experimentation and analyzing the results obtained, we can conclude that in our case is possible to approximate the distribution of evaluations of a real set of texts from a smaller subset of evaluated items. In our experiment, with a test set of 246 items and each text evaluated in a 5-point Likert scale, it would be sufficient to evaluate 24 items (i.e., about 10% of items randomly taken from the entire pool of texts) to ensure that, no matter the prototypical evaluator profile, we obtain a score distribution equivalent to evaluating all the texts generated by the system in at least 90% of the cases. This fact validates the research hypothesis under study: "well-known resampling statistical methods can contribute to get significant results even with a small number of items to be evaluated by each evaluator".

Regarding the already mentioned "n=30 rule-of-thumb" we can say that for the specific experimental setting we achieved good results even without reaching 30 items in the evaluation. Nonetheless, the interesting finding is that for different evaluators and prototypical profiles this number varies and seems that it is not possible to have an ideal number of items for all evaluations beforehand. Considering evaluators with different profiles may mean that approximating the actual distribution of scores requires a higher/lower number of items to be evaluated. Moreover, the minimal number of items depends also on the pre-defined threshold. Thus, the ideal number of

items to obtain reliable results in an NLG evaluation cannot be generalized.

Anyway, our empirical study represents a step forward in the search for an evaluation protocol admitted worldwide. The empirical results highlight the importance of carefully addressing the experimental setting in human evaluation studies for NLG systems. It is crucial to pay special attention to those parameters chosen in the context of the evaluation process, being the number of items especially relevant because it can reduce dramatically the evaluation costs if it is properly selected. In addition, this work provides readers with a benchmark for choosing the ideal number of items for a given evaluation study, since all related resources are available online as open access.[4]

As future work, we plan to extend the empirical study to other types of evaluations in which the scoring criteria and scale may vary from those tested in this work. Also, alternative approaches or formulas for calculating and determining the minimum required number of items to achieve representative results from a sample will be examined. Moreover, we will consider how resampling methods can be integrated in the evaluation procedure to address the lack of resources (e.g., evaluators availability) in NLG human evaluation.

Acknowledgments

J. González-Corbelle, J.M. Alonso-Moral and A. Bugarín-Diz acknowledge the support from the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2019/04 and ED431C2022/19). These grants are co-funded by the European Regional Development Fund (ERDF/FEDER program). In addition, this work is supported by Grants PID2021-123152OB-C21 and PID2020-112623GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future", and by Grant TED2021-130295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR". R.M. Crujeiras acknowledges the support from project PID2020-116587GB-I00, funded by MCIN/AEI/10.13039/501100011033 and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia.

[4] https://gitlab.citius.usc.es/gsi-nlg/human-evaluation-resampling

References

Altman, D. G. 1991. *Practical Statistics for Medical Research*. Chapman and Hall.

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Belz, A. 2022. A Metrological Perspective on Reproducibility in NLP. *Computational Linguistics*, 48(4):1125–1135.

Belz, A., C. Thomson, E. Reiter, and S. Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

De Leon, A. and Y. Zhu. 2008. ANOVA extensions for mixed discrete and continuous data. *Computational Statistics Data Analysis*, 52(4):2218–2227.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41:1149–1160.

Fisher, R. A., 1992. *Breakthroughs in Statistics: Methodology and Distribution*, chapter Statistical Methods for Research Workers, pages 66–70. Springer New York, New York, NY.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

González Corbelle, J., A. Bugarín-Diz, J. Alonso-Moral, and J. Taboada. 2022. Dealing with hallucination and omission in neural natural language generation: A use case on meteorology. In *Proceedings of the 15th International Conference on Natural*

Language Generation, pages 121–130, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Hesterberg, T. 2008. It's time to retire the "n $\geqslant$ 30" rule. In *Proceedings of the American Statistical Association*, Alexandria VA.

Kane, H., M. Y. Kocyigit, A. Abdalla, P. Ajanoh, and M. Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In S. Agarwal, O. Dušek, S. Gehrmann, D. Gkatzia, I. Konstas, E. Van Miltenburg, and S. Santhanam, editors, *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mair, P. and R. Wilcox. 2020. Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, 52:464–488.

Moramarco, F., A. Papadopoulos Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, and A. Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Obeid, J. and E. Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the Transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318, USA. Association for Computational Linguistics.

Reiter, E. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Sellam, T., D. Das, and A. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Student. 1908. Probable error of a correlation coefficient. *Biometrika*, 6(2/3):302–310.

Van der Lee, C., A. Gatt, E. van Miltenburg, and E. Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.1–101151.24.

Wang, J., Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In Y. Dong, W. Xiao, L. Wang, F. Liu, and G. Carenini, editors, *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Hybrid. Association for Computational Linguistics.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*. OpenReview.